DOCUMENT RESUME

ED 339 702 TM 017 313

AUTHOR Gonzalez-Tamayo, Eulogio
TITLE A Closer Look at Test Scores, Selection and Prediction.
PUB DATE 91
NOTE 35p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Comparative Testing; Culture Fair Tests; *Elementary School Students; Foreign Countries; Grade 1; High Achievement; Low Achievement; Minimum Competency Testing; *Predictive Measurement; Predictive Validity; Primary Education; Regression (Statistics); *School Readiness Tests; *Scores; Student Evaluation; *Test Use; Test Validity
IDENTIFIERS *Test of Basic Concepts (Gonzalez Tamayo)

ABSTRACT

Test scores on the Test of Basic Concepts, a school readiness test, were compared with actual performance in school for 186 low-scoring and 133 high-scoring groups of first graders from Spain at the beginning of the school year to explore the issues of competence testing and prediction of performance. The measure of school performance was derived from teacher evaluations at the end of the school year. Data were analyzed in terms of: (1) validity of the decisions based on both validity coefficients and correct classifications; (2) fairness of the classifications; and (3) the meaning and impact of overpredictions and underpredictions. In every analysis, predictor test scores, when used for selection, favored high-scoring groups and individuals. The findings are supported by an analysis of the regression line. Implications of the findings for test use in selecting applicants or for determining competence are discussed. Three tables and one graph present study data. There are 13 references. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Eulogio González-Tamayo

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED339702

A CLOSER LOOK AT TEST SCORES, SELECTION AND PREDICTION

Eulogio González-Tamayo

New York City Public School System

New York, 1991

Running Head: Closer Look

TM017313

BEST COPY AVAILABLE

SUMMARY

Test scores on the Test of Basic Concepts, a school readiness test, were compared with actual performance in school for high and low scoring groups of first graders from Spain. Data were analyzed in terms of (a) validity of the decisions including both validity coefficients and correct classifications, (b) fairness of the classifications and (c) over and underpredictions. In every analysis, predictor test scores, when used for selection, favored high scoring groups and individuals. The findings are supported by an analysis of the regression line.

One of the many uses of psychological tests is as instruments for assessing individuals' present competence in order to make some predictions about their future performance and decide their acceptance or rejection. Competence testing has been used for placing people on programs, for admission to colleges, for selection for jobs, for licensing and relicensing of individuals in education and in other professions, for granting diplomas, etc. In all these situations, the main decision to be made on the basis of test scores is whether to accept or reject the person for the program, the job, the license, etc. There are two assumptions made during the selection: (1) those applicants who are selected on the basis of test scores will perform at an acceptable level in the program, the school, the job or the profession and (2) those applicants who are rejected would have performed unsuccessfully if they have been selected.

The basic problem of competence testing is not different from the problem with tests in general which is the fallibility of test scores in predicting real life performance. This problem, well known to persons involved in testing and selection, has been recently highlighted by the public debate about the devastating and unfair consequences on people's lives when selection decisions are made exclusively on the basis of fallible test scores. Simply put, the fallibility of test scores in selection means that (a) some people selected because they were predicted as successful on the basis of test scores will fail in the school, in college, on the job or in the profession and (b) some persons rejected because they were predicted as unsuccessful on the basis of test scores would have been successful if they have been selected. These are very serious problems affecting the lives of many people and their families.

The problems of competence testing have become so serious that national commissions have been established to study these issues and their

findings raise many questions about the use of test scores in selection. The National Commission on Testing and Public Policy (1990) concluded that "current testing is overrelied upon, lacks adequate public accountability, sometimes leads to unfairness in the allocation of opportunities, and too often undermines vital social policies (p.ix)". The recommendation from the commission was "...alternative forms of assessment must be developed and more critically judged and used, so that testing and assessment open gates of opportunity rather than close them off (p.x)". The Committee on the General Aptitude Test Battery (GATB) (Hartigan and Wigdor, 1989), a test widely used by the Department of Labor to select applicants for jobs in the Federal Government and the foundation of the very well publicized validity generalization theory, concluded that (1) the use of the GATB without some sort of scores adjustment would systematically screen out blacks, some of whom could have performed satisfactory on the job and (2) the inadequacies of the testing technology should not fall more heavily on the social groups already burdened by the effects of past and present discriminations. As a result of this critical report the Department of Labor (DOL) has discontinued its controversial GSTB because of concerns over whether the test adequately serves all individuals. Crouse and Trusheim (1991) reported that as things stand now, it seems that the Scholastic Aptitude Test (SAT) widely used in admission to colleges, is a monument to institutional inertia.

The causes of the problem with tests used in selection were clearly identified in the Report of the National Commission on Testing and Public Policy (1990): (a) there are much larger differences between groups in the predictor test scores than there are in the criterion of performance test scores, and (b) the limited power of tests to predict success in either schools or the work place. The Commission, as Thorndike (1971) had cogently demonstrated twenty years ago, concluded that using test scores

alone to classify people will result in higher rates of misclassification for lower scorers, regardless of whether the test score gap arises from economic and educational disadvantages, from biases in the tests, or from an interaction of these factors.

The conclusions from these national commissions seem to be contrary to the published attitude from the scientific community. A great number of studies have been conducted about the predictive validity of test scores and the regression lines for different groups and the findings have been quite consistent across groups, across tests, and across uses of tests. The findings from these studies have shown (1) aptitude tests are not biased against minority groups (Wigdor & Garner, 1982), at least not against any native-born English speaking minority group (Jensen, 1980, 1984; Hunter et Al., 1984), (2) aptitude tests are equally valid for minority and majority groups (Wigdor & Garner, 1982; Linn, 1978; Boehm, 1977), (3) if there is bias in the test it will favor the minority groups (Jensen, 1980, 1984; Hunter et Al., 1984) and the elimination of bias will result in the selection of few minority persons.

The purpose of the present study is not to solve the problems or to propose a new model of selection but rather to shed some light in the discussion which, unfortunately, has been one sided exaggerating the claims of testing and omitting its weaknesses. The study addresses in the main three issues: (1) validity of the selection based on predictor test scores, (2) fairness of the selection for different groups, and (3) the question of under and overprediction, its meaning and its impact on different groups. To avoid the emotionally laden discussion of the issues, the present study (a) is based on students from Spain rather than from the United States, (b) the data are analyzed by high and low scoring groups rather than by the traditional ethnic or gender groups, and (c) no selection was made and, in that sense, no harm was caused by the use of

test scores.

The general evaluation of a test in terms of its usefulness in selection should be a simple task based on the responses to two basic questions: (a) of those who can perform the job successfully, how many would be selected by the test? the larger this number the more useful the test will be; (b) of those who cannot perform the job successfully, how many are going to be rejected by the test? the larger this number the more useful the test will be. The problem is that, very often, there is no information to answer these questions since test scores are used to select some and to reject other applicants and performance in a real life situation is possible only for those selected but not for the rejected applicants. In some instances, like in licensing, there is not even evaluation of actual performance of those selected. These restrains make any evaluation of the test as an instrument of selection very tentative and unreliable.

Validity of Predictor Scores

The data for the present study include (a) the scores of first graders on the Test of Basic Concepts (González-Tamayo, 1982) used as a indicator of the students' readiness for school work that was administered at the beginning of the school year and (b) the actual school performance as reported by teachers' evaluations at the end of the school year. In order to avoid the confusion between bias of test scores toward some clearly identifiable groups with the issue of misclassifications when test scores are used for selection, the analyses are conducted by high and low scoring groups rather than by sex or any other identifiable personal characteristic. There are males and females in both the high and the low scoring group, and the only criterion for inclusion in one or the other

group was average class score above or below the overall mean. The original groups, school classes, however, were kept intact and all the members of the class, regardless of their scores, were assigned to either the high or the low scoring group. During the following discussion the technical words selection and minimum scores are used when in reality there was no selection and no minimum scores were calculated in the predictor, in order to be selected, nor on the criterion, in order to be classified as successful. The data, however, are real, they could have been but were not used for selection, and the analyses are still proper.

The empirical data provide the opportunity to evaluate the technical soundness of the instrument and also the fairness of the selection if the instrument is used. All the applicants were selected and, therefore, all the applicants were evaluated by the teachers on their performance in the school. This is not the usual case in selection were those accepted and later on evaluated on their performance on the job are not all the applicants but only those selected on the basis of their predictor scores. As a result, some of the applicants will not be selected nor evaluated on their performance, and this group of rejected applicants is important in order to evaluate the usefulness of the test.

Validity Coefficient

The most important indicator of the validity of a selection instrument, no question about it, is the validity coefficient or the correlation between the predictor and the criterion scores. But validity of selection also involves an evaluation of the overall correctness or incorrectness of the classifications of the applicants as result of the selection or rejection. The validity coefficients in the present case were as good as the ones reported in other studies. As all validity coefficients, however, they were modest around 0.50 or lower. This

indicates that only about 25% or less of the observed variations in actual performance can be explained by the variations in the performance on the predictor. Most of the variation in performance, 75%, cannot be explained by variations in the predictor scores. As reported in table 1, the validity coefficient for the low scoring group was only 0.36 while the

Table 1 about here

coefficient for the high scoring group was 0.53. In this respect, the Test of Basic Concepts was somehow a better instrument of selection for the high than for the low scoring group. Both coefficients, however, are within the normal range reported in the literature for the validity coefficients of most of the instruments used in job selection, admission to colleges, placements in the school and professional licensing.

Correctness of the Classifications

Another more appropriate way of evaluating the validity of the test as an instrument of selection would be the one proposed by the APA Standards for Educational and Psychological Tests (1985) "Validity refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores... Validity always refers to the degree to which that evidence supports the inferences that are made from test scores.(p.9)". Test scores are used in selection to classify the applicants into two groups those who score at a minimum predetermined level and will be selected for the job, and those who score below that level and will not be accepted for the job. The procedure makes the inference that those selected will be successful and those rejected would have been unsuccessful on the job if they had been selected. The appropriateness of the inferences will be reflected in the agreement between the inferences and the reality or in the correctness of the

classifications made from the selection instrument.

A correct classification in the present study is one that involves (a) selection of an applicant who will be successful in the school (true positive) or (b) the rejection of applicants who, if selected, would have been unsuccessful in the school (true negatives). A wrong classification is either (a) selection of an applicant who will be unsuccessful in the school (false positives) or (b) rejection of an applicant who would, if selected, have been successful in the school (false negatives). Both true and false classifications must be part of the evaluation of the instrument. It is possible to have one hundred percent of true positives by selecting only the extremely high scorers but, with the selection based on an instrument with less than perfect validity, this probably would also result in high percentage of false negatives or rejection of applicants who would have been successful if selected. The minimum score an applicant should have in order to be selected was arbitrarily set at the mean predictor score from the low scoring group. The minimum score of success was a score of at least 44 on teachers' evaluation; this score represents the predicted score of the minimum predictor score rounded to the nearest number.

As reported in table 2, the selection on the basis of applicants'

Table 2 about here

scores on the Test of Basic Concepts would had produced only a modest 61% (114/186) of correct classifications in the low scoring group and 62% (82/133) in the high scoring group. These figures are a combination of both the true positives, number of applicants selected that were successful, and true negatives, or number of applicants rejected that were unsuccessful in the school. Although the test is more valid for the high (0.53) than for the low (0.36) scoring group, the percentage of correct

classifications is very similar in both cases. An interesting conclusion from this finding is that the validity coefficient does not tell the whole story about the correctness of the inferences made from test scores or about the validity of the test according to the technical standards. If only those applicants scoring at or above the minimum predictor level had been selected, only three out of five applicants would have been correctly classified as either successful or unsuccessful future performers in the school and this figure would be almost identical in both the low and the high scoring groups.

Fairness of the Selection

A good practical definition of fairness or unfairness is provided by the National Commission on Testing and Public Policy (1990): unfairness is the fact that disproportionately fewer minorities who could perform successfully had they been given the opportunity will be selected. While there are other perspectives on fair test use, the performance perspective on fairness will be increasingly important as available human talent is in short supply in the next century. Fairness is always a comparison of the actual selection with actual performance, a fact that is not possible in most selection situations but it was possible in the present case since no student was rejected because of the scores in the predictor. Fairness has nothing to do with comparing predictor scores with predicted scores since it would be redundant to compare predictor scores with other scores calculated from the predictor scores. Later in this report, however, the predicted scores are also analyzed only as an exercise that will clarify some of the technical issues but not as part of the discussion of fairness.

Overall Classifications

Although the number of correct classifications was almost identical in both groups and, as a result, the number of misclassifications were also almost identical in both groups, there was, however, an important distinction between the two groups in the kind of incorrect classifications made although there was no important difference in the total percentage of correct or incorrect classifications. A definition of fairness in this respect would be that the test does not misclassify differently the individuals in both groups. If all the misclassifications in one group are false positives and in the other are all false negatives, although the overall number of misclassification is the same, the test however would not be fair to both groups. In such a case, the group favored by the selection would be the one with the smallest number of false negatives or the lowest number of applicants rejected that would have been successful if they have been selected. The concept of fairness is always in reference to applicants not in reference to the employment institution which obviously will prefer to avoid false positives, selecting individuals that would be unsuccessful, while in terms of applicants the crucial point is to avoid the false negatives.

The total number of misclassifications, about 40% of all the applicants in both groups, included a much higher proportion of false negatives in the low than in the high scoring group. About 30% of all the misclassifications in the low scoring group (22 out of 72) were false negatives or rejected applicants that were successful in school while in the high scoring group this percentage was only about 4% (2 out of 51). This means that many more applicants from the low than from the high scoring group were rejected that could have been successful if they had the opportunity of performing in the school by being selected. The Test of Basic Concepts, if used for selecting applicants who are ready for

school work, would clearly be unfair to the low scoring groups since more potentially successful applicants would be rejected from this than from the high scoring group.

Selection of Successful Applicants

The data in table 2 can be studied by concentrating in the analysis of the columns and the meaning of the different quadrants would be: (a) true positives, or the number of successful applicants that were selected, (b) true negatives, the number of unsuccessful applicants that were rejected, (c) false positive, the number of unsuccessful applicants that were selected and (d) false negatives, the number of successful applicants that were rejected. This analysis will provide some inside into the fate of successful and unsuccessful students if the predictor test had been used as an instrument of selection. An practical definition of fairness in this respect would be represented by (a) selecting the same proportion in both groups of successful applicants and (b) rejecting the same proportion in both groups of unsuccessful applicants. Ideally, all successful applicants should be selected and, therefore, the greater the proportion of successful applicants selected the better the instrument is. Also ideally, none of the unsuccessful applicants should be selected and, therefore, the smaller the number of unsuccessful applicants selected the better the instrument is for the group. A fair test should have the same proportion in both groups of successful applicants selected and of unsuccessful applicants rejected.

From the data in table 2, the following conclusions surface:

1. The number of successful applicants selected, true positives, was 97% for the high scoring group and 76% for the low scoring group. The test in this respect, is not equally fair for both groups and it favors the group with the highest percentage of successful applicants selected,

the high scoring group. The same conclusion could be reached by concentrating the analysis on the false negatives, or the number of successful applicants that were rejected. The numbers also favor the high scoring group since only 3% of the successful applicants were rejected against 24% from the low scoring group. The test clearly is unfair to the low scoring group.

2. The number of unsuccessful applicants rejected, true negatives, was 34% in the high and 47% in the low scoring group. Based on this analysis, the test is not equally fair for both groups and it favors the group with the lowest rate of rejection of unsuccessful applicants or the high scoring group. The same analysis can be explained by saying that more unsuccessful applicants were selected, false positives, from the high (66%) than from the low scoring group (53%). On the basis of this analysis the test is unfair to the low scoring group.

Success of the Selected Applicants

The data in table 2 can also be analyzed by rows and the meaning of the different quadrants would be: (a) true positives, or the number of selected applicants that were successful, (b) true negatives, the number of rejected applicants that were unsuccessful, (c) false positive, the number of selected applicants that were unsuccessful and (d) false negatives, the number of rejected applicants that were successful. The criteria of fairness then would be (a) the same proportion of selected applicants are successful in both groups and (b) the same proportion of rejected applicants are unsuccessful in both groups. The test would be biased in favor of the group that (a) has the lower rate of successful selected applicants since higher number of unsuccessful would be selected and (b) has the higher rate of rejected applicants who were unsuccessful The analysis of the data support the following conclusions

1. A lower proportion of selected applicants were successful, true positives, among the high (54%) than among the low scoring group (58%). The test favors the high scoring group since a lower proportion of the selected applicants were successful than from the low scoring group and as a result more unsuccessful applicants were selected from the high (46%) than from the low scoring group (42%). The analysis indicates that the test is not equally fair to both groups and it favors the group with the highest percentage of unsuccessful selected applicants (high scoring group). There is also another possible interpretation of these results that would make many people uncomfortable. The results could be interpreted as indicating that the test is a better selection instrument for low than for high scoring groups since a higher percentage of the selected applicants would perform better in the low than in the high scoring group. It can also be said that the instrument selected better applicants from the low than from the high scoring group and for this reason more of them succeeded than from the high scoring group. But in terms of fairness, number of people selected that would be successful, the instrument is unfair to the low scoring group.
2. Among the rejected applicants, a greater proportion were unsuccessful in the high (93%) than in the low scoring group (67%). This analysis also indicates that the test favors the high scoring group since in order to be rejected in the high scoring group the applicant has to be unsuccessful while there was a greater likelihood of being rejected in the low scoring group even if the applicant was successful. The reverse of this analysis would be that, among those rejected, a smaller proportion of them were successful in the high (7%) than in the low scoring group (33%). There was a greater probability in the low than in the high scoring group that if the applicant was rejected

he/she could have been successful if selected.

Higher Prediction and Overprediction

Analyses of the regression line have been the bases for the discussion of the prediction of several groups and the general conclusion that the common regression line and the regression line from the high scoring group overpredict the scores of low scoring groups and individuals. Before the data from the present study are presented, it is important to clarify some issues concerning the regression line, its analysis and the conclusions reached on the basis of the information provided by the regression line.

1. Higher prediction of the scores of one group, usually the minority or low scoring group, using the regression line from the other group does not necessarily mean overprediction of the actual performance. It merely means that prediction from the proper line is less advantageous to the low scoring group than the prediction of performance from the common regression line. It is possible that any prædiction, regardless of the regression line used, may be disadvantageous to the minority group. The real test is not overprediction in relation to the proper regression line but overprediction in relation to actual performance on the job or in the school. This point will discussed in a greater detail later since it represents the pivotal point of this part of the study.
2. The regression line is good for a group as a whole, but it does not necessarily mean that it is good for individuals specifically. The regression line takes into account the raw individual scores but it does not analyze the scores in the pass-fail dichotomy as necessary in

order to understand the misclassifications and correct classifications made by using the predictor or the predicted scores as the basis of selection. Correct classifications is the issue, not predicted scores. It is possible that the regression line makes a correct prediction for the group as an average of the predicted scores, but it makes many misclassifications. A classification is correct when the applicant (a) is selected and is successful on the job or (b) is rejected and is unsuccessful on the job.

3. The best prediction regression line merely means that overall, counting the over and the underpredictions made from the regression line, the result is a minimum amount of error. This, again, does not mean that the regression line makes the least amount of error of classification.
4. The equation representing the regression line is based on the assumption of a bivariate normal distribution of scores. There are several criterion scores for each ability level and several ability scores for each criterion score. The regression line passes through the middle of these distributions of scores. This clearly indicates that, at each ability level, there are always under and overprediction of performance scores but in the group, these over and underpredictions cancel each other for a minimum amount of error in overall prediction.

It is possible that the over and underpredictions at each level of ability may not cancel each other, but overall, the regression line may still be the best fit regression line or the line that overall will produce the least amount of errors in the prediction. Additionally, the selection is not done through the entire regression line but only on the part representing scores above a predetermined minimum score. It is possible that at each ability level the average error in prediction is zero but the number of misclassifications is positive. In the series of performance scores such as 2,3,5,6,15,24 for a particular level of

ability, the average predicted score is 8; regardless of the minimum level of performance, the average error will always be zero. The number of misclassifications, however, if the minimum level of performance is the average, will be four misclassifications that will be predicted as successful but in actuality were unsuccessful because their actual performance was below the minimum (8).

In the present study, the regression line from the low scoring group was used in the analysis as the prediction line for all the applicants because it had a higher intercept than the regression line for the high scoring group and, as an average, they perform better in the school than the high scoring group. The average performance, as reported in table 1, was 43.60 for the low scoring group and 42.26 for the high scoring group. As a result, if the regression line of the low scoring group is used for predicting the scores of the other group, most of the predicted scores of the high scoring group will be overpredicted as it is represented clearly in Figure 1. An analysis of the regression line would suggest that many

Figure 1 about here

more criterion scores will be overpredicted in the high than in the low scoring group.

The results of the comparison between the predicted and the actual performance scores are reported in table 3. The overall analysis of the

Table 3 about here

predictions indicates that about half of the predicted performance scores were overpredicted in both the low and the high scoring groups. About 52% of the predicted scores in the high and 43% in the low scoring group were higher than the actual performance scores of the applicants in the school as evaluated by the teachers. There was a greater number of

overpredictions in the high than in the low scoring group.

This finding clearly demonstrates that the regression line does not consistently over or underpredicts the scores but rather that the under and the overpredictions are very similar in both groups. The finding seems to be contrary to the general reported analysis of the regression line and also contrary to an analysis of Figure 1. It has been consistently reported, and Figure 1 demonstrates, that using the regression line of the higher performance group, the prediction will favor the low performance scoring group. This is not the case since both groups have similar percentages of over and underpredicted criterion score. One explanation for this apparent contradiction is that most reported results have been based on the theoretical analysis of the regression line not on the comparison between the predicted scores from the regression line and the actual criterion scores as done in the present study.

Prediction and Selection

An analysis of the columns in table 3 will provide information about the origin of the over and underpredictions. Which proportion of the over and underpredictions is from the selected and which is from the group of rejected applicants? The analysis of the data in the columns will provide the answer to this basic question.

1. An analysis of the underpredicted scores column indicates that, in the high scoring group, a greater number of underpredicted scores were among the rejected (58%) than among the selected applicants (42%). The opposite was the case with the low scoring group: a greater number of underpredictions occurred among the selected (64%) than among the rejected applicants (36%). The significant of this finding is that a substantial number of underpredictions in both groups are from the selected and from the rejected applicants. A theoretical analysis of

the regression line would indicate that most of the underpredictions take place among the rejected applicants or those scoring below the minimum predictor score. The empirical data from the present study do not support that conclusion.

2. The column of overpredictions indicates that a larger proportion of the overpredictions in the high scoring group were among the applicants rejected (69%),and the reverse was the case in the low scoring groups, a larger proportion of the overpredictions were among the selected (64%) than among the rejected applicants. As in the underpredictions, the overpredictions occurred in any score and in both groups but not consistently among applicants in one group nor consistently among selected or rejected applicants.
3. It is important to note that a low or high predictor score is no indication that the predicted criterion score will be under or overpredicted. The empirical data do not support the conclusion from the analysis of the regression line that scores above a point will overpredict criterion scores and scores below that point will underpredict the criterion scores. Under and overpredictions occur all along the continuum because real data are not as neat as the theoretical case may indicate. There is clearly a distinction between higher prediction of the average predictor score and overprediction of the individual scores at each level of the predictor score.

Selection and Prediction

An analysis of the rows of the data in table 3 will provide information about the fate of the prediction of criterion scores, in terms of being under or overpredicted, for both the group of selected and the group of rejected applicants. The analysis in the previous section provided information about where the over and underpredictions were coming

from, from the selected or from the rejected group; the present analysis is about the predictions for the selected and the rejected groups of applicants. The two questions are not the same and the answers to both of them is necessary in order to better understand the problem of over and underprediction.

1. There were substantial number of overpredictions and underpredictions among both the selected and the rejected applicants and in both the high and the low scoring group. The over or underpredictions do not show any exclusivity for either the selected or rejected applicants or for the high or low scoring groups. All the groups are substantially represented in both the over and the underpredictions.
2. Among the selected applicants the amount of overprediction was substantial and very similar in both the high (46%) and the low scoring group (43%). A selected applicant in either the low or the high scoring group has similar probability of overpredicted test scores (one out of every two).
3. Among the rejected applicants, there was a substantial difference in overprediction between the high (74%) and the low scoring group (43%) but the amount of overprediction was substantial in both groups. It is important to note that the low scoring group line was used to calculate the predictions and as a result it was expected that the high scoring group will have more overpredictions than the low scoring group. This result is consistent with the reported results and with the analysis of the prediction line.
4. Many rejected applicants from the high scoring group (26%) had unpredicted scores which is contrary to the conclusion from figure 1 where all the rejected applicants, lower part of the line, should have overpredicted scores. The finding, however, is consistent with the analysis of the regression line performed earlier where it was

demonstrated that the prediction is for the average score but, in reality, at each level there are both under and overpredicitons.

Discussion of the Findings

One of the biggest jokes in testing has been the outrageous statement that tests used in selection favor minorities or low scoring groups (Jensen, 1980, 1984; Hunter et Al., 1984). The present findings do not support that claim, on the contrary, if there is a general conclusion it is that, whatever way the data are analyzed, it is clear that decisions based on test scores favor the higher scoring individuals and groups. The specific findings of the present study are as follows.

1. The validity coefficient of a test used in selection, predictive validity, does not tell the whole story about the validity of test scores. A better indicator would be the correctness of the decisions made on the basis of test scores, including both correct and incorrect decisions, or correct classifications and misclassification. The test of Basic Concepts, if used to select children who are prepared for school work, will produce a modest 61% of correct classifications in both the high and the low scoring groups, although the validity coefficients were different (0.36 in the low and 0.53 in the high scoring group). It seems that the test would be equally valid for both groups of students in terms of the correct classifications made.
2. An analysis of the misclassifications, about 40% in either group, reveals (a) a larger proportion of them in the low (30%) than in the high scoring group (4%) were in the false negative decisions indicating that a larger proportion of students from the low than from the high scoring group were rejected who could have been successful if selected;

(b) a natural consequence of the previous analysis, if the total number of misclassifications is the same in both groups, is that a larger proportion of the misclassifications in the high (96%) than in the low scoring group (70%) were false positive decisions indicating than many more students from the high than from the low scoring group were accepted who in reality were unsuccessful. In both counts, accepting unsuccessful and rejecting successful applicants, the TBC is not equally fair to both groups and in both instances it favors the high scoring group.

3. A comparison of successful applicants from both groups indicates that (a) much larger proportion of them were selected in the high (97%) than in the low scoring group (76%); (b) the other side of the same decision indicates that many more successful applicants were rejected from the low (24%) than from the high scoring group (3%). In both counts, accepting and rejecting successful applicants, the TBC was not equally fair to both groups and in both instances it favored the high scoring group: more of the successful applicants from this group were selected than from the low scoring group but higher number of the successful applicants were rejected from the low than from the high scoring group.
4. A comparison of unsuccessful applicants indicates that (a) a larger proportion of them were selected in the high (66%) than from the low scoring group (53%); (b) a smaller proportion of unsuccessful applicants were rejected in the high (34%) than in the low scoring group (47%). In whatever way you look to these figures, the message is clear and the same as previous analyses: the TBC clearly favors the high scoring group by selecting a higher proportion of unsuccessful applicants than from the low scoring group and by rejecting a smaller proportion of successful applicants.
5. A comparison of the fate of the selected applicants indicate that (a)

a very slight higher proportion of them were successful in the low (58%) than in the high scoring group (54%) and (b) a slighter higher proportion of the selected applicants was unsuccessful in the high (46%) than in the low scoring group (42%). The conclusion from these findings is that although there is not a clear leaning of the selection in favor of one or the other group, the selection on the basis of predictor test scores does not favor the low scoring group.

6. A comparison of rejected applicants indicates that (a) a much larger proportion of them were successful in the low (24%) than in the high scoring group (3%) and (b) smaller proportion of the rejected applicants was unsuccessful in the low (67%) than in the high scoring group (93%). The story keeps repeating itself: the TBC favors the high scoring group by rejecting a greater proportion of unsuccessful applicants and a smaller proportion of successful applicants than in the low scoring group.
7. A comparison of the prediction with actual performance on the job indicates where the over or underprediction, if any, occurred. The analyses indicate that (a) there was a higher number of underpredictions in the low (57%) than in the high scoring group (48%) and (b) there were substantial numbers of under and overprediction in both groups, very similar numbers. Both of these findings are contrary to the reported findings from a theoretical analysis of the regression line . Using the regression line from the high scoring group, theoretically, the low scoring group will have a greater proportion of overpredictions than underpredictions and a greater proportion of overpredictions than in the high scoring group. None of these theoretical predictions materialized in the present study.

The purpose of the present study was to shed some light in the discussion about the fairness issues of test scores used in selection,

by addressing the following three important issues: (1) validity of the selection based on predictor test scores, (2) fairness of the selection for different groups, and (3) the question of under and overprediction, its meaning and its impact of different groups. With respect to validity, it is clear than it involves more than just the predictive validity coefficient and the analysis of the classifications provides insightful information. With respect to the fairness of test scores in relation to low and high scoring groups it is clear that the test scores are not fair to the low scoring group and they clearly favor the high scoring group. Whatever way you analyze the data the conclusion is a litany of "in favor of the high scoring group". In ten different analyses of the data, not a single one supported the conclusion that test scores, if used to select applicants who are ready for school, will favor the low scoring group and all of them clearly indicate that test scores favor the high scoring group. Finally, the over and underprediction is a statistical concept and it is not very useful in practice.

Theoretical Support for The Conclusions

To put into perspective the present and many other findings in selection, it is important to realize three important psychometric facts about the prediction which are very well known among tests experts but which have not been widely publicized.

1. The regression line, like any other statistic in the classical theory of tests, is group dependent and it has to be analyzed very carefully when comparing groups that are different in the ability being measured by the predictor. The regression equations for two different groups (Mh =/ Ml), each one of them having the same means and the same

standard deviations in both the predictor and the criterion, are given by

$$Yh = r.X + (1 - r)Mh$$
$$Yl = r.X + (1 - r)Ml$$

where r is the validity of the test and Mh and Ml represent the mean of the high and the low scoring group respectively.

If the validity is perfect (r = 1), then the predicted criterion scores are identical to the predictor scores and they are the same in the high and in the low scoring groups; an individual will obtain the same predicted criterion score regardless of group membership. But if the validity is less than one, group membership becomes a factor in determining the predicted criterion score an individual will get and group membership becomes more relevant depending on (a) how low the validity coefficient is and (b) how large the group differences are. If the validity is 0.60, an individual score of 90 will predict a criterion score of 94 if the group mean is 100, but only 87.6 if the group mean is 84; the same case but when the validity is 0.30 will produce a predicted criterion score of 97 if the group mean is 100 and only 85.8 if the group mean is 84. These numbers make many people uncomfortable because it is against common sense, even though it is psychometrically correct, that the same person can have two different predicted scores depending on the group to which, by accident or by nature, the person belongs.

2. If the predicted criterion score is used for selection instead of the predictor score, an error is made in the prediction whenever the validity is less than one, and this error of prediction is the difference between actual and predicted performance. The correlation of this error with actual performance in school or on the job (González-Tamayo, 1985) is given by

$$R_{ey} = \sqrt{1 - R^2_{xy}}$$

where Rxy represents the validity of the test, and the regression of error of prediction over actual criterion scores would be

$$e = (1 - R^2_{xy})(Y - My)$$

The error of prediction for an individual is (1) indirectly related to the validity of the test and the lower the validity the larger the error of any particular score and (2) is directly related to the departure of the actual criterion score from the group mean and the error of prediction will be positive for performance scores above the mean and negative for performance score below the mean. The net effect of these positive and negative errors throughout the entire distribution is a minimum and the regression line does what it is required to do, that is, to minimize the errors of prediction. However, when the individual scores are considered, the error of prediction favors the high scoring group and individuals. The high scorers predicted performance scores have positive error of prediction (are inflated) and the low scoring individuals predicted performance scores have negative errors of prediction (are deflated). The actual difference in predicted performance between low and high scorers individuals and groups are augmented twice by inflating the predicted scores of high scorers and deflating the predicted scores of the low scorers. The invalidity of the test clearly favors the high scoring individuals and groups.

3. A similar analysis can be done comparing the scores on the predictor and the actual performance and the result of this analysis will also be that the predictor scores favor high scoring individuals and groups and the advantage is larger the lower the validity of the test is. Using

the predictor instead of the actual performance for the selection some error is introduced because the correlation between predictor and performance scores is never perfect. This error is called the error of selection to differentiate it from the error of prediction, and it would be given by

$$e = (1 - Rxy)(X - Mx)$$

This error is positively related to the predictor scores and the errors are positive for score above the mean and negative for scores below the mean. The difference in predictor scores between high and low scoring individuals and groups is increased twice because, under the condition of minimizing the errors through the least square fit mathematical formula, always inflates the true scores of the high scorers and deflates the true scores of the low scorers. Contrary to what Jensen (1980) reported, an analysis of the error made in selection leads to the conclusion that invalidity in a test favors high scoring individuals and groups.

Tests are a product of human endeavor and they, probably more than anything else, reflect the limitations of human nature. People are good judges only of mediocre performance, the one that we observed most frequently, but they are not so good in evaluating the extreme performances, those which are very good or very poor. Human judgments outside the ordinary, middle performance, make exaggerations and a good performance becomes terrific and a bad performance becomes horrendous. Tests are very similar to any other human evaluation: they are very good at assessing the middle of any ability, they exaggerate good and bad performance. It would be good for testing if testing experts pay attention to the recommendation from The Committee on the General Aptitude

Test Battery (GATB) (Hartigan and Wigdor, 1989): " (1) the claims for the testing programs be tempered and that employers as well as job seekers be given a balanced view of the strengths and weaknesses of the GATB and (2) officials refrain from making any dollar estimates of the gains that would result from test-based selection (p.286)".

BIBLIOGRAPHY

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1985). Standards for Educational and Psychological Testing. Washington D.C.: Author.

Boehm, V.R. (1977). Differential Prediction: A Methodological Artifact? Journal of Applied Psychology, 62, 146-154.

Crouse, J. and Trusheim, D. (1991). How Colleges Can Correctly Determine Selection Benefits from the SAT. Harvard Educational Review, 61(2), 125-147.

González-Tamayo, E. (1982). Test de Conceptos Elementales. Madrid: Instituto Calasanz de Ciencias de la Educación (ICCE).

González-Tamayo, E. (1985). Bias in Prediction: A Closer Look at Theory, Findings, and Conclusions. ERIC Document Reproduction Service, No. Ed. 254-534.

Hartigan, J. A. and Wigdor, A. K. (Eds.). (1989). Fairness in Employment Testing: Validity Generalization, Minority Issues, and the General Aptitude Test Battery (GATB). Committee on the GATB Commission on Behavioral and Social Sciences and Education. National Research Council, National Academy Press: Washington, D. C.

Hunter, J. E., Schmidt, F. L. and Rauschenbeger, J. (1984). Methodological, Statistical, and Ethical Issues in the Study of Bias in Psychological Tests. In Reynolds, C. R. and Brown, R. T. (Eds.), Perspectives on Bias in Mental Testing. Plenum Press: New York.

Jensen, A.R. (1980). Bias in Mental Testing. New York: The Free Press.

Jensen, A.R. (1984). Test Bias: Concepts and Criticisms. In C.R. Reynolds and R.T. Brown (Eds.) Perspectives on Bias in Mental

Testing. New York: Plenum Press.

Linn, R.L. (1978). Single-group Validity, Differential Validity, and Differential Prediction. Journal of Applied Psychology, 63, 507-512.

National Commission on Testing and public Policy. (1990). From Gatekeeper to Gateway: Transforming Testing in America. Author. Boston College, Boston, MA.

Thorndike, R.L. (1971). Concepts of Culture-Fairness. Journal of Educational Measurement, 8, 63-70.

Wigdor, A. K. and Garner, W. R. (Eds.). (1982). Ability Testing: Uses, Consequences, and Controversies. National Academy Press, Washington, D. C.

Figure 1: Graphic Representation of the Prediction Equations of Low and High Scoring Groups in the Test of Basic Concepts

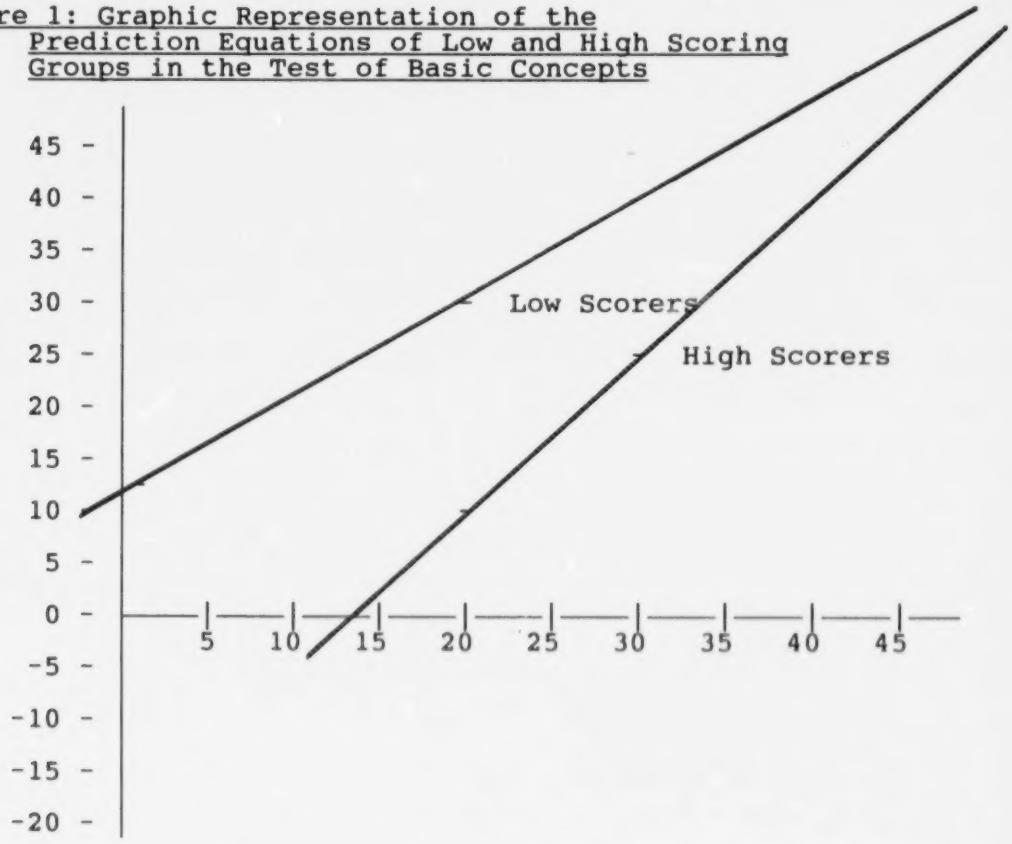

Table 1: Means and Standard Deviations in the Basic Concepts Test and in School Performance and the Correlation Coefficients

	Low Scorers			High Scorers		
	N	Mean	S. D.	N	Mean	S. D.
TBC	186	38.71	4.57	133	40.84	4.15
School Per.	186	43.60	9.86	133	42.26	11.80
Correlation		0.36			0.53	

Table 2: Outcomes of Selection Decisions for High and Low Scorers on the Basis of the Test of Basic Concepts and Actual School Performance

LOW SCORERS

PREDICTOR	CRITERION: FAIL		CRITERION: PASS		CRITERION: TOTAL	
PASS	50 0.42	0.53	69 0.58	0.76	119 1.00	0.64
FAIL	45 0.67	0.47	22 0.33	0.24	67 1.00	0.36
TOTAL	95 0.51	1.00	91 0.49	1.00	186 1.00	1.00

HIGH SCORERS

PREDICTOR	CRITERION: FAIL		CRITERION: PASS		CRITERION: TOTAL	
PASS	49 0.46	0.66	57 0.54	0.97	106 1.00	0.80
FAIL	25 0.93	0.34	2 0.07	0.03	27 1.00	0.20
TOTAL	74 0.56	1.00	59 0.44	1.00	133 1.00	1.00

Table 3: Comparison of Predicted and Actual School Performance Using the Regression Line with the Highest Intercept (Low Scoring Group)

LOW SCORERS

PREDICTOR	OVERPREDICTED		UNDERPREDICTED		TOTAL	
SELECTED	51 0.43	0.64	68 0.57	0.64	119 1.00	0.64
REJECTED	29 0.43	0.36	38 0.57	0.36	67 1.00	0.36
TOTAL	80 0.43	1.00	106 0.57	1.00	186 1.00	1.00

HIGH SCORERS

PREDICTOR	OVERPREDICTED		UNDERPREDICTED		TOTAL	
SELECTED	49 0.46	0.31	57 0.54	0.42	106 1.00	0.80
REJECTED	20 0.74	0.69	16 0.26	0.58	55 1.00	0.20
TOTAL	69 0.52	1.00	64 0.48	1.00	133 1.00	1.00